Глава 8

Аппроксимация нейронными сетями

Данная глава посвящена вопросам представления и аппроксимации функций с помощью нейронных сетей. Основное внимание будет уделено равномерной аппроксимации непрерывных функций, заданных на компакте, «широкими» нейронными сетями, принадлежащими классам вида $\mathcal{F}_{+,\sigma,m}$. В конце главы будет исследован вопрос влияния глубины нейронных сетей на их аппроксимирующие свойства.

8.1 Предварительные рассуждения

С практической точки зрения особый интерес представляют два типа функций. Это булевы функции и непрерывные вещественнозначные функции, заданные на некотором компакте в $\mathbb{R}^m$ ($m \in \mathbb{N}$).

Булевы функции представляют собой отображения вида $\{0,1\}^m \longrightarrow \{0,1\}$, которые, как это хорошо известно (см. [64]), могут быть заданы формулами при помощи операций конъюнкции $\wedge$, дизъюнкции $\vee$ и отрицания $^-$. Таким образом, если данные операции могут быть выражены нейронными сетями, то и любая формула может быть трансформирована в нейронную сеть, реализующую соответствующую ей булеву функцию.

Следующие определения

$$\begin{aligned} h_\vee(v_1, v_2) &\coloneqq \sigma_{01}(v_1 + v_2 - 0.5), \\ h_\wedge(v_1, v_2) &\coloneqq \sigma_{01}(v_1 + v_2 - 1.5), \\ h_-(v_1) &\coloneqq \sigma_{01}(-v_1 + 0.5) \quad \big(v_1, v_2 \in \{0,1\}\big), \end{aligned}$$

показывают, что рассматриваемые операции могут выражены через нейронные сети с σ_{01}-узлами. Конъюнкции выражается через $h_\wedge$, дизъюнкции через $h_\vee$ и отрицания через h_-. Следовательно, любая булева функция может быть реализована нейронной сетью с σ_{01}-узлами.

Пример 8.1. В качестве примера рассмотрим булеву функцию

$$f(v_1, v_2) = (\overline{v}_1 \wedge v_2) \vee (v_1 \wedge \overline{v}_2), \tag{8.1}$$

называемую *операцией исключающего «или»*. На кортежах $(0,0)$ и $(1,1)$ она принимает значение 0, а на кортежах $(0,1)$ и $(1,0)$ она принимает значение 1.

Интересно отметить, что эти два набора кортежей не могут быть линейно отделены друг от друга. Поэтому функция f не может быть выражена одним σ_{01}-нейроном.

На рис. 8.1 изображена нейронная сеть, реализующая функцию f. Эта сеть построена в соответствии со структурой формулы в правой части (8.1). Заметим, что существует нейронная сеть, содержащая всего три σ_{01}-узла, которая также реализует функцию f.

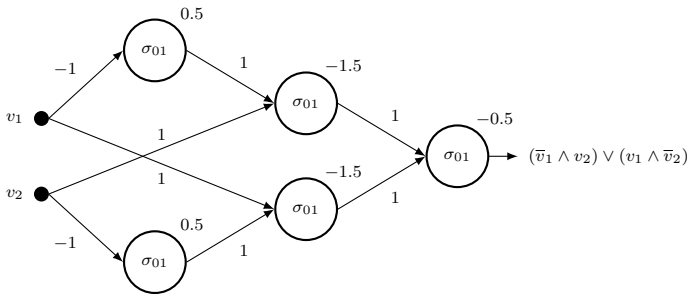

Рис. 8.1: Нейронная сеть, реализующая операцию исключающего «или».

Перейдём к рассмотрению случая непрерывных функций. Начнём с знаменитого результата о представлении непрерывной функции в виде суперпозиции «простых» функций.

Теорема 8.1 (Колмогоров [65])**.** *Любая непрерывная функция $f \in C(I_m)$ $(m \geqslant 2)$ может быть представлена в виде*

$$f(u_1, \ldots, u_m) = \sum_{j=1}^{2m+1} \chi_j \Big(\sum_{i=1}^{m} \psi_{ij}(u_j) \Big),$$

где χ_j, ψ_{ij} суть непрерывные функции одного переменного. При этом функции ψ_{ij} монотонны и не зависят от f.

Пример 8.2. В соответствии с теоремой 8.1 функция двух переменных будет иметь представление

$$\begin{aligned} f(u_1, u_2) =& \chi_1\big(\psi_{11}(u_1) + \psi_{21}(u_2)\big) + \chi_2\big(\psi_{12}(u_1) + \psi_{22}(u_2)\big) \\ &+ \ldots + \chi_5\big(\psi_{15}(u_1) + \psi_{25}(u_2)\big). \end{aligned}$$

Фактически теорема Колмогорова говорит о возможности точного представления непрерывной функции 3-х слойной нейронной сетью. Правда эта нейронная сеть будет иметь очень специфические функции активации, некоторые из которых зависят от самой представляемой функции. В случае использования общеупотребительных функций активации, скорее всего, можно говорить только о приближении нейронными сетями.

Далее, будет рассматриваться задача аппроксимации непрерывных функций, заданных на m-мерном единичном кубе I_m $(m \in \mathbb{N})$. При этом, соответствующие результаты могут быть обобщены на произвольное компактное множество в $\mathbb{R}^m$.

8.2 Элементарные методы

Начнём с рассмотрения частных случаев. Покажем, что липшицевые функции одного переменного равномерно аппроксимируются нейронными сетями из класса $\mathcal{F}_{+,\sigma_{01},1}$.

Определение 8.1. Функция $f : \mathbb{R}^m \longrightarrow \mathbb{R}$ $(m \in \mathbb{N})$ называется *α-липшицевой* $(\alpha > 0)$, если

$$|f(\mathbf{u}_1) - f(\mathbf{u}_2)| \leqslant \alpha \, \|\mathbf{u}_1 - \mathbf{u}_2\|_\infty \quad (\mathbf{u}_1, \mathbf{u}_2 \in \mathbb{R}^m).$$

Теорема 8.2. *Пусть $g : \mathbb{R} \longrightarrow \mathbb{R}$ является α-липшицевой функцией для некоторого $\alpha > 0$.*

Тогда для любого $\varepsilon > 0$ существует 2-х слойная нейронная сеть h, содержащая $\lceil \frac{\alpha}{\varepsilon} \rceil$ σ_{01}-узлов, такая, что

$$\sup_{u \in [0,1]} |g(u) - h(u)| \leqslant \varepsilon. \tag{8.2}$$

◄ Зафиксируем произвольное $\varepsilon > 0$. Обозначим $m \coloneqq \lceil \frac{\alpha}{\varepsilon} \rceil$ и по определению положим

$$h \coloneqq \sum_{i=1}^{m} a_i \mathbf{1}_{u \geqslant b_i},$$

где

$$b_i \coloneqq \frac{(i-1)\varepsilon}{\alpha} \quad (i = 1, \ldots, m), \quad a_1 \coloneqq g(b_1),\ a_i \coloneqq g(b_i) - g(b_{i-1}) \quad (i = 2, \ldots, m).$$

Зафиксируем произвольное $u \in [0, 1]$. Обозначим через k максимальный индекс такой, что $u \geqslant b_k$. Заметим, что на отрезке $[b_k, u]$ функция h постоянна.

Следовательно,

$$\begin{aligned} |g(u) - h(u)| &\leqslant |g(u) - g(b_k)| + |g(b_k) - h(b_k)| + |h(b_k) - h(u)| \\ &\leqslant \alpha |u - b_k| + \left| g(b_k) - \sum_{i=1}^{k} a_i \right| + 0 \\ &\leqslant \alpha \frac{\varepsilon}{\alpha} + \left| g(b_k) - g(b_1) - \sum_{i=2}^{k} \big(g(b_i) - g(b_{i-1})\big) \right| \\ &= \varepsilon, \end{aligned}$$

а значит выполняется искомое неравенство (8.2). ■

Разберём случай аппроксимации липшицевых функций многих переменных нейронными сетями с σ_{relu}-узлами. В начале докажем вспомогательное утверждение.

Утверждение 8.1. *Пусть* $R \coloneqq [b_1, c_1) \times \ldots \times [b_m, c_m) \subseteq I_m$ $(m \in \mathbb{N})$.

Тогда для любого $\varepsilon > 0$ *существует 2-х слойная нейронная сеть* h*, состоящая из* $4m + 1$ σ_{relu}*-узлов такая, что*

$$\|h - \mathbf{1}_R\|_{\mathcal{L}^1(I_m)} < \varepsilon. \tag{8.3}$$

◄ Для каждого $\gamma > 0$ и индекса i $(i = 1, \ldots, m)$ определим функцию

$$\begin{aligned} h_{\gamma,i}(u) \coloneqq \sigma_{\mathrm{relu}}\left(\frac{u - (b_i - \gamma)}{\gamma}\right) - \sigma_{\mathrm{relu}}\left(\frac{u - b_i}{\gamma}\right) \\ - \sigma_{\mathrm{relu}}\left(\frac{u - c_i}{\gamma}\right) + \sigma_{\mathrm{relu}}\left(\frac{u - (c_i + \gamma)}{\gamma}\right) \qquad (u \in \mathbb{R}). \end{aligned}$$

Эта функция обладает следующим свойством

$$h_{\gamma,i}(u) \in \begin{cases} \{1\}, & \text{если} \quad u \in [b_i, c_i]; \\ \{0\}, & \text{если} \quad u \notin [b_i - \gamma, c_i + \gamma]; \\ [0, 1], & \text{иначе.} \end{cases}$$

Построим нейронную сеть h_γ, состоящую из $4m + 1$ нейрона с функцией активации σ_{relu}. По определению положим

$$h_\gamma(\mathbf{u}) \coloneqq \sigma_{\mathrm{relu}}\left(\sum_{i=1}^{m} h_{\gamma,i}(u) - (m-1)\right) \qquad (\mathbf{u} \in \mathbb{R}^m).$$

Эта нейронная сеть обладает следующим свойством

$$h_\gamma(\mathbf{u}) \in \begin{cases} \{1\}, & \text{если} \quad \mathbf{u} \in R; \\ \{0\}, & \text{если} \quad \mathbf{u} \notin R_\gamma; \\ [0, 1], & \text{иначе,} \end{cases}$$

где $R_\gamma \coloneqq [b_1 - \gamma, c_1 + \gamma) \times \ldots \times [b_m - \gamma, c_m + \gamma)$.

Запишем цепочку неравенств

$$\begin{aligned}\|h_\gamma - \mathbf{1}_R\|_{\mathcal{L}^1(I_m)} &\leqslant \|h_\gamma - \mathbf{1}_R\|_{\mathcal{L}^1(\mathbb{R}^m)} \\ &= \int\limits_R |h_\gamma - \mathbf{1}_R|\, d\lambda_m + \int\limits_{R_\gamma \setminus R} |h_\gamma - \mathbf{1}_R|\, d\lambda_m + \int\limits_{\mathbb{R}^m \setminus R_\gamma} |h_\gamma - \mathbf{1}_R|\, d\lambda_m \\ &= \int\limits_{R_\gamma \setminus R} |h_\gamma - \mathbf{1}_R|\, d\lambda_m \\ &\leqslant \prod_{i=1}^m (c_i - b_i + 2\gamma) - \prod_{i=1}^m (c_i - b_i) \longrightarrow 0 \quad \text{при} \quad \gamma \longrightarrow +0.\end{aligned}$$

Это означает существование нейронной сети, удовлетворяющей требуемому неравенству (8.3). ■

Теорема 8.3. *Пусть $g : \mathbb{R}^m \longrightarrow \mathbb{R}$ $(m \in \mathbb{N})$ является α-липшицевой функцией для некоторого $\alpha > 0$.*

Тогда для любого $\varepsilon > 0$ существует 3-х слойная нейронная сеть h, содержащая не более чем $(4m+1)\left\lceil \frac{2\alpha}{\varepsilon} \right\rceil^m$ σ_{relu}-узлов такая, что

$$\|h - g\|_{\mathcal{L}^1(I_m)} < \varepsilon. \tag{8.4}$$

◀ Зафиксируем произвольное $\varepsilon > 0$ и обозначим

$$d \coloneqq \left\lceil \frac{2\alpha}{\varepsilon} \right\rceil, \quad M \coloneqq d^m.$$

Разобьём множество $[0,1)^m$ на M попарно непересекающихся полуоткрытых кубов, каждая сторона которых имеет длину $1/d$. Запишем

$$[0,1)^m = \bigcup_{j=1}^M R_j, \qquad R_j \coloneqq \left[b_1^{(j)}, c_1^{(j)}\right) \times \ldots \times \left[b_m^{(j)}, c_m^{(j)}\right).$$

По построению

$$c_i^{(j)} - b_i^{(j)} = \frac{1}{d} \qquad (i = 1, \ldots, m;\ j = 1, \ldots, M).$$

Введём обозначения

$$a_j \coloneqq g\left(\mathbf{b}^{(j)}\right) \quad (j = 1, \ldots, M), \qquad A \coloneqq |a_1| + \ldots + |a_M|.$$

Определим вспомогательную функцию

$$f \coloneqq \sum_{j=1}^M a_j \mathbf{1}_{R_j}.$$

Для любого вектора $\mathbf{u} \in [0,1)^m$ существует уникальный индекс $j_{\mathbf{u}}$ такой, что $\mathbf{u} \in \mathbf{1}_{R_{j_{\mathbf{u}}}}$. Следовательно,

$$|f(\mathbf{u}) - g(\mathbf{u})| = |a_{j_{\mathbf{u}}} - g(\mathbf{u})| \leqslant \alpha \|\mathbf{u} - \mathbf{b}^{(j_{\mathbf{u}})}\|_\infty \leqslant \frac{\alpha}{d} \leqslant \frac{\varepsilon}{2}. \tag{8.5}$$

Далее, возможны два случая.

Случай 1, $A = 0$. Это означает, что функция f принимает только нулевое значение. Учитывая неравенство (8.5), в качестве искомой нейронной сети возьмём $\sigma_{\mathrm{relu}}(\langle \mathbf{0}, \cdot \rangle)$.

Случай 2, $A \neq 0$. В силу утв. 8.1 существуют 2-х слойные нейронные сети $h_1, \ldots, h_M$, каждая из которых состоит из $4m+1$ σ_{relu}-узлов, такие, что

$$\|h_j - \mathbf{1}_{R_j}\|_{\mathcal{L}^1(I_m)} < \frac{\varepsilon}{2A} \qquad (j = 1, \ldots, M). \tag{8.6}$$

По определению положим

$$h \coloneqq \sum_{j=1}^{M} a_j h_j.$$

Заметим, что функция h представляет собой 3-х слойную нейронную сеть, состоящую из $(4m+1)M^m$ σ_{relu}-узлов.

Окончательно получим

$$\begin{aligned}
\|h - g\|_{\mathcal{L}^1(I_m)} &\leqslant \|h - f\|_{\mathcal{L}^1(I_m)} + \|h - f\|_{\mathcal{L}^1(I_m)} \leqslant \Big| \ {\scriptstyle (8.5)} \ \Big| \\
&\leqslant \|h - f\|_{\mathcal{L}^1(I_m)} + \frac{\varepsilon}{2} \\
&\leqslant \sum_{j=1}^{M} |a_j| \|h_j - \mathbf{1}_{R_j}\|_{\mathcal{L}^1(I_m)} + \frac{\varepsilon}{2} \leqslant \Big| \ {\scriptstyle (8.6)} \ \Big| \\
&< \sum_{j=1}^{M} |a_j| \frac{\varepsilon}{2A} + \frac{\varepsilon}{2} \\
&= \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.
\end{aligned}$$

■

Отметим некоторые ограничения полученных результатов, которые предстоит преодолеть в следующих разделах. Вместо класса непрерывных функций задача аппроксимации решалась для более узкого класса липшицевых функций. В многомерном случае вместо равномерного приближения строилось приближение по интегральной норме. Утверждения были доказаны для конкретных функций активации.

8.3 Равномерная аппроксимация

В данном разделе будет показано, что непрерывная функция может быть равномерно приближена 2-х слойной нейронной сетью, использующей во внутреннем слое сигмоидную функцию активации. Сформулируем основной результат.

Теорема 8.4 (Цыбенко [66])**.** *Пусть σ – сигмоидная функция. Тогда класс нейронных сетей $\mathcal{F}_{+,\sigma,m}$ является всюду плотным в $C(I_m)$ $(m \in \mathbb{N})$.*

В начале сформулируем и получим ряд вспомогательных результатов (см. [6, 49]).

Теорема 8.5 (Рисс-Марков-Какутани)**.** *Предположим, что на множестве Ω задана структура хаусдорфова и компактного метрического пространства.*

Тогда для любого непрерывного линейного функционала $F \in C(\Omega)^*$ существует единственная регулярная мера со знаком ν_F такая, что

$$F(f) = \int_{\Omega} f(\omega) \nu_F(d\omega) \qquad \big(f \in C(\Omega)\big).$$

Утверждение 8.2. *Пусть* $(\mathsf{L}, \|\cdot\|)$ – *нормированное пространство,* $\mathsf{L}' \subseteq \mathsf{L}$ – *линейное подпространство,* $x_0 \in \mathsf{L} \setminus \mathsf{L}'$ *и* $\delta > 0$. *Предположим, что* $\|x_0 - x\| \geqslant \delta$ *для всех* $x \in \mathsf{L}'$.

Тогда существует непрерывный линейный функционал $F : \mathsf{L} \longrightarrow \mathbb{R}$ *такой, что*

- $\|F\| \leqslant 1$;
- $F(x) = 0$ *для всех* $x \in \mathsf{L}'$;
- $F(x_0) = \delta$.

◀ На линейном подпространстве

$$\mathsf{L}'' \coloneqq \big\{x + \lambda x_0 \,:\, x \in \mathsf{L}', \lambda \in \mathbb{R}\big\} \subseteq \mathsf{L}$$

определим функцию $F : \mathsf{L}'' \longrightarrow \mathbb{R}$, полагая

$$F(x + \lambda x_0) \coloneqq \lambda\delta \qquad \big(x \in \mathsf{L}', \lambda \in \mathbb{R}\big).$$

Для любых $x_1, x_2, x \in \mathsf{L}'$ и $\lambda_1, \lambda_2, \lambda, \alpha \in \mathbb{R}$ выполняются равенства

$$\begin{aligned} F((x_1 + \lambda_1 x_0) + (x_2 + \lambda_2 x_0)) &= F(x_1 + x_2 + (\lambda_1 + \lambda_2)x_0) \\ &= (\lambda_1 + \lambda_2)\delta \\ &= F(x_1 + \lambda_1 x_0) + F(x_2 + \lambda_2 x_0); \\ F(\alpha(x + \lambda x_0)) &= \alpha\lambda\delta \\ &= \alpha F(x + \lambda x_0). \end{aligned}$$

Таким образом, F является линейным функционалом на L''. По построению

$$\begin{aligned} F(x_0) &= F(0 + 1 \cdot x_0) = 1 \cdot \delta = \delta, \\ F(x) &= F(x + 0 \cdot x_0) = 0 \cdot \delta = 0 \qquad \big(x \in \mathsf{L}'\big). \end{aligned}$$

Проверим выполнение неравенства

$$F(x + \lambda x_0) \leqslant \|x + \lambda x_0\| \qquad \big(x \in \mathsf{L}', \lambda \in \mathbb{R}\big).$$

При $\lambda = 0$ это неравенство очевидно выполняется. Поэтому будем предполагать $\lambda \neq 0$. По предположению

$$\Big\|x_0 + \frac{x}{\lambda}\Big\| \geqslant \delta,$$

а значит $|\lambda|\delta \leqslant \|x + \lambda x_0\|$. Следовательно,

$$F(x + \lambda x_0) = \lambda\delta \leqslant |\lambda|\delta \leqslant \|x + \lambda x_0\|.$$

По теореме Хана-Банаха F может быть продолжен до линейного функционала, заданного на всём пространстве L, для которого будет выполняться неравенство

$$F(y) \leqslant \|y\| \qquad \big(y \in \mathsf{L}\big).$$

Это означает, что F является непрерывным линейным функционалом и $\|F\| \leqslant 1$. ■

Доказанное утверждение может быть переформулировано в следующем виде.

Утверждение 8.3. *Пусть* $(\mathsf{L}, \|\cdot\|)$ – *нормированное пространство и* $\mathsf{L}' \subseteq \mathsf{L}$ – *не всюду плотное в* L *линейное подпространство. Тогда существует непрерывный линейный функционал* F, *заданный на* L, *такой, что* $F \neq 0$ *и* $F|_{\mathsf{L}'} = 0$.

Вернёмся к непосредственному доказательству теоремы Цыбенко.

Утверждение 8.4. *Пусть $\mathcal{F}$ – линейное не всюду плотное подпространство в $C(I_m)$ $(m \in \mathbb{N})$. Тогда существует регулярная мера $\mu \neq 0$ такая, что*

$$\int\limits_{I_m} f d\mu = 0 \qquad \big(\textit{для всех } f \in \mathcal{F}\big).$$

◀ В силу утв. 8.3 существует непрерывный линейный функционал $F : C(I_m) \longrightarrow \mathbb{R}$ такой, что $F \neq 0$ и $F|_{\mathcal{F}} = 0$.

Для завершения доказательства достаточно заметить, что в силу теоремы Рисса-Маркова-Какутани существует ненулевая регулярная мера со знаком μ такая, что

$$F(g) = \int\limits_{I_m} g d\mu \qquad \big(g \in C(I_m)\big).$$

■

Лемма 8.1. *Пусть σ – дискриминационная функция. Тогда класс нейронных сетей $\mathcal{F}_{+,\sigma,m}$ является всюду плотным в $C(I_m)$ $(m \in \mathbb{N})$.*

◀ Предположим обратное. Класс функций $\mathcal{F}_{+,\sigma,m}$ является линейным подпространством в $C(I_m)$. Согласно утв. 8.4 найдётся ненулевая регулярная мера $\mu \neq 0$ такая, что

$$\int\limits_{I_m} f d\mu = 0 \qquad \big(\text{для всех } f \in \mathcal{F}_{+,\sigma,m}\big).$$

Но это входит в противоречие с предположением о том, что функция σ является дискриминационной. ■

◀ (доказательство теоремы 8.4) Теорема 7.1 говорит о том, что непрерывная функция является дискриминационной.

Применяя лемму 8.1, получим доказательство теоремы. ■

8.4 Равномерная аппроксимация (продолжение)

Как показывает следующий пример, теорема Цыбенко может использоваться для установления аппроксимирующих свойств класса нейронных сетей вида $\mathcal{F}_{+,\sigma,m}$ даже в случае, когда функция активации σ не является сигмоидной. Для этого необходимо с помощью σ надлежащим образом сконструировать сигмоидную функцию. Точнее показать, что класс $\mathcal{F}_{+,\sigma,1}$ содержит сигмоидную функцию.

Пример 8.3. Функция

$$\sigma(u) \coloneqq \sigma_{\text{relu}}(u) - \sigma_{\text{relu}}(u-1) \qquad (u \in \mathbb{R}) \tag{8.7}$$

является сигмоидной. Поэтому в силу теоремы Цыбенко класс $\mathcal{F}_{+,\sigma,m}$ всюду плотен в $C(I_m)$ $(m \in \mathbb{N})$.

Любая нейронная сеть N такая, что $h_N \in \mathcal{F}_{+,\sigma,m}$, может быть модифицирована следующим образом. Руководствуясь правилом (8.7), каждый σ-узел сети N может быть заменён на два σ_{relu}-узла. В результате будет получена новая нейронная сеть N' такая, что $h_N = h_{N'} \in \mathcal{F}_{+,\sigma_{\text{relu}},m}$.

Следовательно, $\mathcal{F}_{+,\sigma,m} \subseteq \mathcal{F}_{+,\sigma_{\text{relu}},m}$, а значит класс $\mathcal{F}_{+,\sigma_{\text{relu}},m}$ всюду плотен в $C(I_m)$.

Существует ряд независимо полученных доказательств теоремы Цыбенко. Особого внимания заслуживает доказательство из работы [67], в основу которого положена теорема Стоуна-Вейерштрасса (см. [6]). Обсудим основные идеи этого доказательства.

Определение 8.2. Пусть K – хаусдорфовый компакт. Множество непрерывных функкий $C_\circ \subseteq C(K)$ называется *алгеброй Стоуна*, если одновременно выполняются следующие условия:

1. если $f, g \in C_\circ$ и $c \in \mathbb{R}$, то $cf, f+g, fg \in C_\circ$;

2. постоянная функция $1 \in C_\circ$;

3. для любых различных $x_1, x_2 \in K$ существует $f \in C_\circ$ такая, что $f(x_1) \neq f(x_2)$.

Теорема 8.6 (Стоун-Вейерштрасс)**.** *Пусть K – хаусдорфовый компакт и $C_\circ \subseteq C(K)$ – алгебра Стоуна. Тогда $C_\circ$ является всюду плотным в $C(K)$.*

Приведём два примера использования теоремы Стоуна-Вейерштрасса для доказательства аппроксимирующих свойств классов функций.

Пример 8.4. Покажем, что класс $\mathcal{F}_{+,\cos,m}$ является всюду плотным в $C(I_m)$ $(m \in \mathbb{N})$. Для этого достаточно установить, что он является алгеброй Стоуна.

Проверим первое свойство. Прежде всего заметим, что $\mathcal{F}_{+,\cos,m}$ является линейным пространством. Далее, используя известное тождество

$$\cos(u_1)\cos(u_2) = \frac{1}{2}\big[\cos(u_1+u_2) + \cos(u_1-u_2)\big] \qquad (u_1, u_2 \in \mathbb{R})$$

получим

$$\Bigg[\sum_{j=1}^{M} a_j \cos\big(\langle \mathbf{w}_j, \mathbf{u}\rangle + b_j\big)\Bigg] \cdot \Bigg[\sum_{k=1}^{M'} a'_k \cos\big(\langle \mathbf{w}'_k, \mathbf{u}\rangle + b'_k\big)\Bigg] =$$

$$\frac{1}{2}\sum_{j=1}^{M}\sum_{k=1}^{M'} a_j a'_k \Big[\cos\big(\langle \mathbf{w}_j + \mathbf{w}'_k, \mathbf{u}\rangle + b_j + b'_k\big) + \cos\big(\langle \mathbf{w}_j - \mathbf{w}'_k, \mathbf{u}\rangle + b_j - b'_k\big)\Big]$$

$$(\mathbf{w}_j, \mathbf{w}'_k, \mathbf{u} \in \mathbb{R}^m;\ a_j, a'_k, b_j, b'_k \in \mathbb{R}),$$

а это в свою очередь означает, что класс $\mathcal{F}_{+,\cos,m}$ замкнут относительно операции произведения своих элементов.

Второе свойство следует из представления

$$1 = \cos\big(\langle \mathbf{0}, \mathbf{u}\rangle\big) \qquad (\mathbf{u} \in \mathbb{R}^m).$$

Проверим третье свойство. Выберем произвольные $\mathbf{u}', \mathbf{u}'' \in \mathbb{R}^m$ $(\mathbf{u}' \neq \mathbf{u}'')$ и определим функцию

$$f(\mathbf{u}) \coloneqq \cos\big(\langle \mathbf{u} - \mathbf{u}'', \mathbf{u}' - \mathbf{u}''\rangle / \|\mathbf{u}' - \mathbf{u}''\|_2^2\big) \qquad (\mathbf{u} \in \mathbb{R}^m).$$

Эта функция принадлежит классу $\mathcal{F}_{+,\cos,m}$ и

$$f(\mathbf{u}') = 1 \neq 0 = f(\mathbf{u}'').$$

Пример 8.5. Покажем, что класс $\mathcal{F}_{+,\exp,m}$ является всюду плотным в $C(I_m)$ $(m \in \mathbb{N})$. Для этого достаточно установить, что он является алгеброй Стоуна.

Проверим первое свойство. Прежде всего заметим, что $\mathcal{F}_{+,\exp,m}$ является линейным пространством. Далее, получим

$$\left[\sum_{j=1}^{M} a_j \exp\left(\langle \mathbf{w}_j, \mathbf{u}\rangle + b_j\right)\right] \cdot \left[\sum_{k=1}^{M'} a'_k \exp\left(\langle \mathbf{w}'_k, \mathbf{u}\rangle + b'_k\right)\right] =$$

$$\sum_{j=1}^{M}\sum_{k=1}^{M'} a_j a'_k \exp\left(\langle \mathbf{w}_j + \mathbf{w}'_k, \mathbf{u}\rangle + b_j + b'_k\right)$$

$$(\mathbf{w}_j, \mathbf{w}'_k, \mathbf{u} \in \mathbb{R}^m;\ a_j, a'_k, b_j, b'_k \in \mathbb{R}),$$

Второе свойство следует из представления

$$1 = \exp\left(\langle \mathbf{0}, \mathbf{u}\rangle\right) \qquad (\mathbf{u} \in \mathbb{R}^m).$$

Проверим третье свойство. Выберем произвольные $\mathbf{u}', \mathbf{u}'' \in \mathbb{R}^m$ $(\mathbf{u}' \neq \mathbf{u}'')$ и определим функцию

$$f(\mathbf{u}) \coloneqq \exp\left(\langle \mathbf{u} - \mathbf{u}'', \mathbf{u}' - \mathbf{u}''\rangle / \|\mathbf{u}' - \mathbf{u}''\|_2^2\right) \qquad (\mathbf{u} \in \mathbb{R}^m).$$

Эта функция принадлежит классу $\mathcal{F}_{+,\exp,m}$ и

$$f(\mathbf{u}') = e \neq 1 = f(\mathbf{u}'').$$

Возвращаясь к работе [67], отметим, что приводимое в ней доказательство состоит из двух частей. Первая часть фактически была разобрана в двух рассмотренных выше примерах. Во второй части доказывается равномерная аппроксимируемость функций $\exp$ $(\cos)$ элементами класса $\mathcal{F}_{+,\sigma,1}$ с сигмоидной функцией σ.

Окончательный ответ об аппроксимирующих свойствах классов нейронных сетей вида $\mathcal{F}_{+,\sigma,m}$ с непрерывными функциями активации даёт следующая теорема, которая здесь приводится без доказательства.

Теорема 8.7 (Лешно и др. [68])**.** *Пусть $\sigma : \mathbb{R} \longrightarrow \mathbb{R}$ – непрерывная функция, не являющаяся полиномом. Тогда класс нейронных сетей $\mathcal{F}_{+,\sigma,m}$ является всюду плотным в $C(I_m)$ $(m \in \mathbb{N})$.*

Заметим, что отказаться от условия, требующего, чтобы функция активации не была бы полиномом, нельзя. Если предположить, что она является полиномом степени k, то построить адекватные равномерные аппроксимации для полиномов степени $k+1$ уже не получится.

8.5 Глубокие сети

Результаты, полученные в предыдущих разделах, говорят о том, что задача аппроксимации непрерывной функции может быть решена при промощи построения «неглубокой» нейронной сети. Возникает закономерный вопрос. Существуют ли какие-либо преимущества за счёт использования именно глубоких нейронных сетей? Оказывается, что глубокие нейронные сети могут иметь существенно меньшую сложность, где под сложностью понимается количество используемых узлов.

Теорема 8.8 (Телгарский [69,70])**.** *Для любого $L \geqslant 2$ существует функция $g \in C([0,1])$, которая реализуется нейронной сетью из класса $\mathcal{F}_{+,\sigma_{\mathrm{relu}},1}^{L^2+2}$ с $2L^2+4$ σ_{relu}-узлами, и обладает следующим свойством.*

Для любой функции $h \in C([0,1])$, которая реализуется нейронной сетью из класса $\mathcal{F}_{+,\sigma_{\mathrm{relu}},1}^{L'}$ $(L' \leqslant L)$ с числом σ_{relu}-узлов, не превосходящим 2^L, выполняется неравенство

$$\|g - h\|_{\mathcal{L}^1([0,1])} \geqslant \frac{1}{32}. \tag{8.8}$$

Доказательство проведём для частного случая $L' = 1$.

Вспомогательные построения

Нас будут интересовать свойства суперпозиций следующей функции

$$\Lambda(u) \coloneqq \begin{cases} 2u, & \text{если} \quad u \in \left[0, \frac{1}{2}\right); \\ 2 - 2u, & \text{если} \quad u \in \left[\frac{1}{2}, 1\right]; \\ 0, & \text{иначе.} \end{cases}$$

Можно заметить, что график функции Λ^l ($l \in \mathbb{N}$) на отрезке $[0, 1]$ обладает следующими свойствами (рис. 8.2). Часть графика на отрезке $[1/2, 1]$ получается путём зеркального отражения части графика на отрезке $[0, 1/2]$ относительно прямой $u = 1/2$. При $l \geqslant 2$ график функции Λ^l на отрезке $[0, 1/2]$ получается из графика функции Λ^{l-1} на отрезке $[0, 1]$ путём сжатия и параллельного переноса последнего.

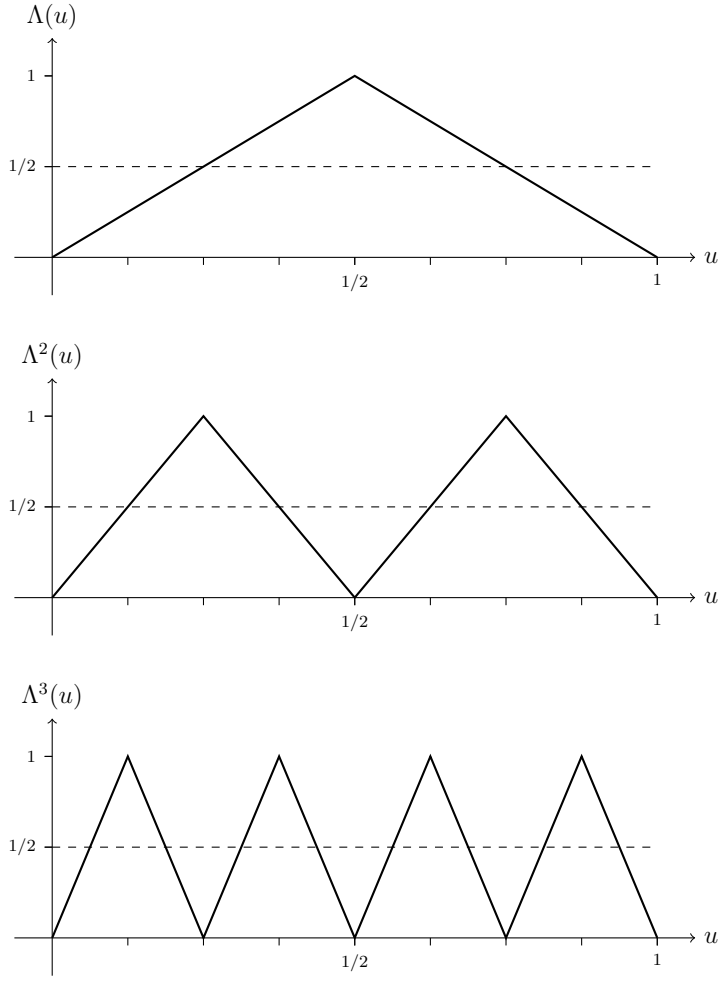

Рис. 8.2: Графики функций Λ, Λ^2, Λ^3.

График функции Λ^l и прямая $u \mapsto 1/2$ определяют систему треугольников на плоскости. Одна сторона такого треугольника, которую будем называть *основанием*, лежит на прямой $u \mapsto 1/2$. Противоположная основанию вершина треугольника лежит либо на прямой $u \mapsto 1$, либо на прямой $u \mapsto 0$. В первом случае треугольник будем называть *верхним*, а во втором случае – *нижним*.

Общее число определённых таким образом треугольников равно

$$C_\Lambda(l) \coloneqq 2^l - 1 \qquad (l \in \mathbb{N}),$$

при этом каждый треугольник будет иметь площадь

$$S_\Lambda(l) \coloneqq 2^{-l-2} \qquad (l \in \mathbb{N}).$$

Определение 8.3. Пусть $h \in C([0,1])$ и T – треугольник, порождённый графиком функции Λ^l $(l \in \mathbb{N})$ и прямой $u \mapsto 1/2$. Будем называть этот треугольник *h-выжившим*, если существует отрезок $[\alpha, \beta] \subset [0,1]$ такой, что

- проекция основания треугольника T на прямую $u \mapsto 0$ содержится в $[\alpha, \beta]$ (будем также говорить, что треугольник принадлежит этому отрезку);
- либо T является верхним треугольником и $h(u) \leqslant \frac{1}{2}$ для всех $u \in [\alpha, \beta]$, либо T является нижним треугольником и $h(u) \geqslant \frac{1}{2}$ для всех $u \in [\alpha, \beta]$.

Число всех h-выживших треугольников будем обозначать через $C_h(l)$.

Заметим, что число верхних и число нижних треугольников, принадлежащих любому отрезку, могут отличаться не больше, чем на единицу.

Доказательство теоремы Телгарского

Простые геометрические соображения позволяют получить следующую оценку.

Утверждение 8.5. *Пусть $h \in C([0,1])$. Тогда*

$$\|\Lambda^l - h\|_{\mathcal{L}^1([0,1])} \geqslant S_\Lambda(l)\, C_h(l) = 2^{-l-2}\, C_h(l) \qquad (l \in \mathbb{N}).$$

Утверждение 8.6. *Пусть $h \in C([0,1])$. Предположим, что существует разбиение отрезка*

$$0 = \alpha_1 < \alpha_2 < \ldots < \alpha_K < \alpha_{K+1} = 1 \qquad (K \in \mathbb{N}), \tag{8.9}$$

обладающее следующим свойством. Для любого k $(1 \leqslant k \leqslant K)$ либо $h(u) \leqslant \frac{1}{2}$ для всех $u \in [\alpha_k, \alpha_{k+1}]$, либо $h(u) \geqslant \frac{1}{2}$ для всех $u \in [\alpha_k, \alpha_{k+1}]$.

Тогда

$$C_h(l) \geqslant \frac{1}{2}\big[C_\Lambda(l) - 2K + 1\big] = \frac{1}{2}\big[2^l - 2K\big] \qquad (l \in \mathbb{N}). \tag{8.10}$$

◀ Всего график функции Λ^l и прямая $u \mapsto 1/2$ определяют $C_\Lambda(l)$ треугольников. Если проекция основания треугольника на прямую $u \mapsto 0$ содержит некоторый узел разбиения α_k $(2 \leqslant k \leqslant K)$, то такой треугольник может не являться h-выжившим. Число таких треугольников не превосходит $K - 1$. Отбросим их.

Здесь стоит заметить, что крайние узлы α_1 и α_{K+1} не могут попасть в проекцию основания ни одного из треугольников.

Каждый из оставшихся треугольников будет принадлежать одному из отрезков $[\alpha_k, \alpha_{k+1}]$ $(1 \leqslant k \leqslant K)$.

Как уже отмечалось ранее, общее число верхних и общее число нижних треугольников, принадлежащих любому из отрезков $[\alpha_k, \alpha_{k+1}]$, могут отличаться не больше, чем на единицу. Поэтому отбросив ещё не более чем K треугольников, мы получим множество, половину которого составляют h-выжившие треугольники. Отсюда и вытекает справедливость оценки (8.10). ■

Утверждение 8.7. *Пусть $h \in C([0,1])$ – кусочно-линейная функция и M – общее количество точек на отрезке $[0,1]$, в которых нарушается линейность функции h. Тогда*

$$C_h(l) \geqslant \frac{1}{2}\big[2^l - 2(M+2)\big] \qquad (l \in \mathbb{N}). \tag{8.11}$$

◄ Воспользуемся разбиением (8.9) отрезка $[0,1]$ для функции h из предыдущего утв. 8.6. Предположим, что значение параметра K является наименьшим из всех возможных. Это означает, что отсутствуют два подряд идущих отрезка разбиения, на которых функция h одновременно больше (меньше) либо равна $1/2$.

В этом случае каждый отрезок разбиения будет содержать либо точку, в которой нарушается линейность функции h, либо концевую точку разбиения α_1 или α_{K+1}. Следовательно, $K \leqslant M+2$.

Объединяя это неравенство и неравенство (8.10) из предыдущего утв. 8.6, получим требуемую оценку (8.11). ■

Пример 8.6. Рассмотрим функцию h, график которой изображен на рис. 8.3. Линейность этой функции нарушается в двух точках. Для функции Λ^4 имеется семь h-выживших треугольников.

С учётом сказанного выше, оценка (8.11) из утв. 8.7 примет вид

$$7 = C_h(4) \geqslant \frac{1}{2}\big[2^4 - 2(2+2)\big] = 4.$$

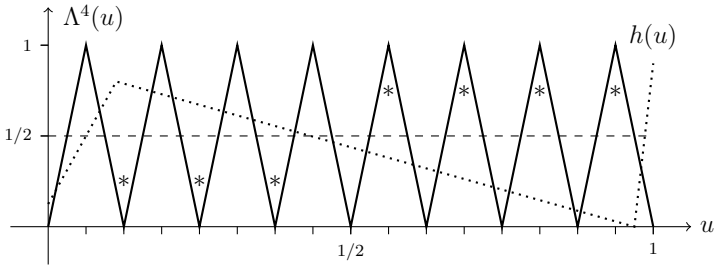

Рис. 8.3: Примеры h-выживших треугольников (выделены звёздочкой).

Утверждение 8.8. *Функция Λ^l ($l \in \mathbb{N}$) может быть реализована нейронной сетью из класса $\mathcal{F}^l_{+,\sigma_{\mathrm{relu}},1}$, содержащей $2l$ σ_{relu}-узла.*

◄ Для функции Λ требуемая нейронная сеть определяется равенством

$$\Lambda(u) = \sigma_{\mathrm{relu}}(2u) - \sigma_{\mathrm{relu}}(4u-2) \qquad (u \in \mathbb{R}).$$

На рис. 8.4 на примере функции Λ^2 показано как из нейронной сети, реализующей функцию Λ^{l-1}, путём добавления одного нового скрытого слоя с двумя σ_{relu}-узлами получается нейронная сеть, реализующей функцию Λ^l. При этом, если первая нейронная сеть удовлетворяет требуемому условию, то этому условию будет удовлетворять и вторая нейронная сеть. ■

◄ (доказательство теоремы 8.8) Положим $g \coloneqq \Lambda^{L^2+2}$. В силу утв. 8.8 выбранная таким образом функция g будет удовлетворять условию теоремы.

Зафиксируем произвольную функцию h, которая реализуется нейронной сетью из класса $\mathcal{F}_{+,\sigma_{\mathrm{relu}},1}$ с числом σ_{relu}-узлов, не превосходящим 2^L, Эта функция является кусочно-линейной и общее количество точек, в которых нарушается её линейность, $M \leqslant 2^L$.

Из утв. 8.7 следует, что

$$C_h(L^2+2) \geqslant \frac{1}{2}\big[2^{L^2+2} - 2^{L+1} - 4\big].$$

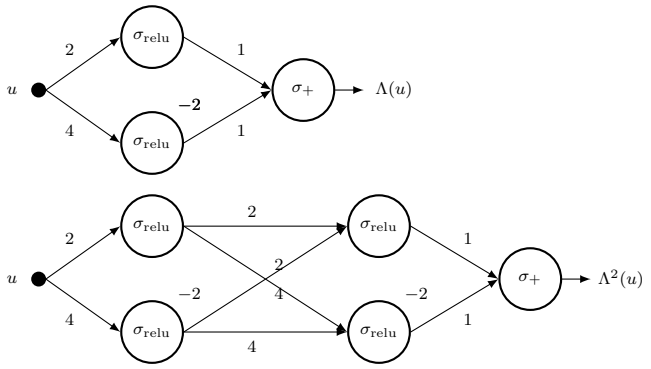

Рис. 8.4: Нейронные сети, реализующие функции Λ и Λ^2.

Применяя утв. 8.5, получим

$$\begin{aligned}\|g-h\|_{\mathcal{L}^1([0,1])} &\geqslant 2^{-L^2-4}\, C_h(L^2+2)\\ &\geqslant \frac{1}{32}\left[4-2^{-L^2+L+1}-2^{-L^2+2}\right]\\ &\geqslant \frac{1}{32}.\end{aligned}$$

■